

Multi-Level Feature Descriptor for Robust Texture Classification via Locality-Constrained Collaborative Strategy

Shu Kong and Donghui Wang
 {aimerykong, dhwang}@zju.edu.cn

College of Computer Science and Technology, Zhejiang University
 Hangzhou, China

February 24, 2013

Abstract

This paper introduces a simple but highly efficient ensemble for robust texture classification, which can effectively deal with translation, scale and changes of significant viewpoint problems. The proposed method first inherits the spirit of spatial pyramid matching model (SPM), which is popular for encoding spatial distribution of local features, but in a flexible way, partitioning the original image into different levels and incorporating different overlapping patterns of each level. This flexible setup helps capture the informative features and produces sufficient local feature codes by some well-chosen aggregation statistics or pooling operations within each partitioned region, even when only a few sample images are available for training. Then each texture image is represented by several orderless feature codes and thereby all the training data form a reliable feature pond. Finally, to take full advantage of this feature pond, we develop a collaborative representation-based strategy with locality constraint (LC-CRC) for the final classification, and experimental results on three well-known public texture datasets demonstrate the proposed approach is very competitive and even outperforms several state-of-the-art methods. Particularly, when only a few samples of each category are available for training, our approach still achieves very high classification performance.

1 Introduction

Texture is widely considered as a fundamental ingredient of the structure of natural images, and texture classification is an important problem in computer vision with many applications. Yet despite almost 50 years of research and development, designing a high-accuracy and robust texture classification system for real-world applications remains a challenge for at least three reasons: the wide range of various natural texture types; the presence of large intra-class variations in texture images, such as rotation, scale, viewpoint, and even non-rigid surface deformation, caused by arbitrary viewing and illumination conditions; and the demands of low computational complexity and a desire to limit algorithm tuning [2].

There are four basic elements that constitute a reliable texture classification system, as Liu *et al.* point out in [3]: (1) local texture descriptors, (2) non-local statistical descriptors, (3) the design of a distance/similarity measure, and (4) the choice of classifier. Thanks to the emergence of *Bag-of-Feature* words (BoF) model, which treats an image as a collection of unordered appearance descriptors extracted from local patches, quantizes them into discrete "visual words" and then computes a compact histogram representation for semantic image classification. As a result, recent

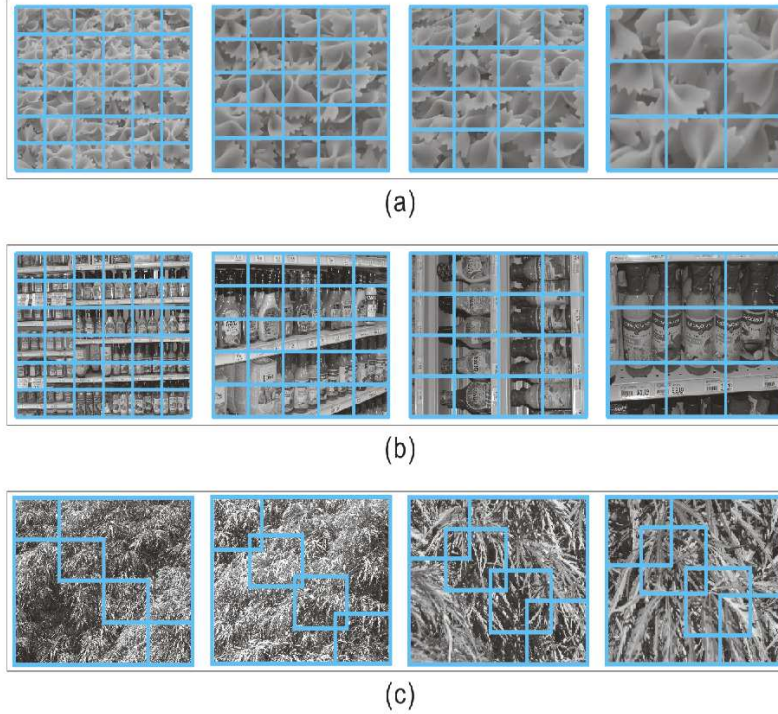


Figure 1: Several samples of three categories from UMD texture database [1]. From (a) and (b), we can see the statistic information within of the regions from various partition levels can capture multiple-scale feature information. As a result, local scale and translation differences can be effectively alleviated. (c) presents different overlapping patterns in 4×4 partition. Here only the diagonal regions are plotted for better illustration. This overlapping partition strategy helps us capture reliable and redundant information of the textures.

interest for texture classification tends to represent a texture non-locally by the distribution of local textons [4, 5, 6, 2], and achieves state-of-the-art performance.

As an extension of BoF, *spatial pyramid matching* model (SPM) [7] has emerged as a popular framework to represent an image by extracting image descriptors such as SIFT [8] or HOG [9] on a dense grid, encoding them over a learned dictionary, and then summarizing the distribution of the codes in the cells of a spatial pyramid by some well-chosen aggregation statistics, or pooling operation. SPM paradigm has made a remarkable success on a range of image classification benchmarks, and becomes a major component of the state-of-the-art systems [10, 11, 12]. Inspired by SPM, we introduce a similar framework to SPM to partition an image into increasingly fine segments, but in a more flexible way, exploiting multi-level partitions with various overlapping patterns and thereby forming redundant local texture feature codes for each regions by a pooling operation. In this way, our method produces a reliable feature pond containing these informative feature codes, even when only a few samples of each class are available for training.

To take full advantage of the feature pond, we develop a simple but effective and efficient mechanism for the final classification, called *collaborative representation-based classification with locality constraint*, LC-CRC for short, which is similar in appearance to *sparse representation-based classification* (SRC) [13], but essentially differs in two ways: (1) ℓ_2 -norm regularization is adopted in the

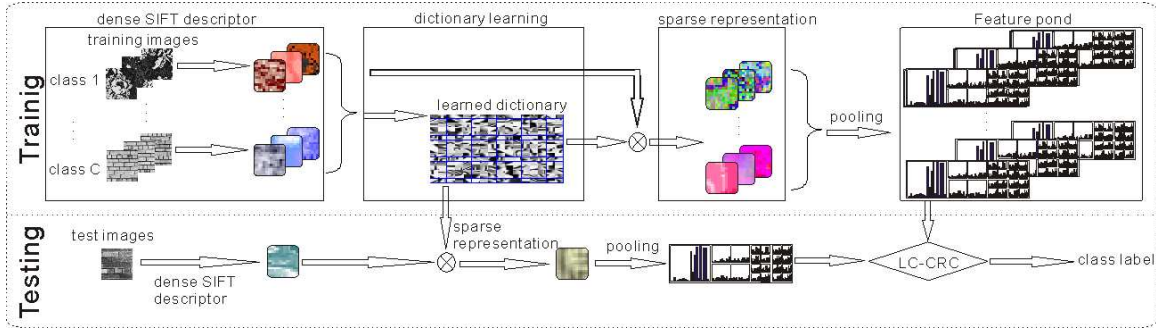


Figure 2: The flowchart of our proposed robust texture classification approach.

least square fitting problem rather than ℓ_1 -norm penalty, and (2) locality constraint is employed to speed up classification process.

We summarize the advantages of our texture classification system below:

- Different from many state-of-the-art texture classification methods which combine several types of descriptors, our approach uses only a single type of feature descriptor, *i.e.* SIFT descriptor [8]. Thus, our method is much simple but still much capable of discriminating textures demonstrated in the experiment.
- Benefiting from the flexible partition strategy, the proposed method can produce redundant feature codes to form a reliable feature pond, even though only a few samples of each category are available for training.
- Instead of the widely-used SVMs, a simple but effective classification mechanism, LC-CRC, is developed in our method. It facilitates overall translation, scale, and viewpoint invariance in classification. And experimental results demonstrate the suggested LC-CRC is very effective and efficient.

The rest of this paper is organized as follows. Section 2 gives a very brief review on the related work. Details of our framework are elaborated in Section 3. We show experimental results in Section 4 and conclude with some discussion in Section 5 before closing.

2 Related Work

We review closely related work on spatial pyramid matching (SPM), sparse representation-based classification paradigm (SRC), and local coordinate coding (LCC).

SPM framework has made a remarkable success on a range of image classification benchmarks, and remains a major component of the state-of-the-art systems [14, 7, 15, 11, 12]. In SPM, spatial order of local descriptors is seriously considered in image classification tasks, however, it is of little importance and captures no essential features in texture classification, because texture can be coined as "a subtle balance between repetition and innovation" [16], or approximative reduplication. Despite of this, SPM indeed inspires us to borrow the idea of multi-level scheme (pyramid partition of the image) to design a local invariant framework for texture classification. This is one focus of our work — adopting a more flexible partition configuration with multiple overlapping patterns such

as $\{3, 4, 5\}^1$, as Figure 1 shows, in place of the common fashion of SPM, *e.g.* $\{1, 2, 4\}$ in [10] and $\{1, 2, 4, 8\}$ in [12], .

Despite the widely-used SVM classifier, *sparse representation-based classification* (SRC) [13] achieves a great success in face recognition task, and it boosts the research of sparsity based pattern classification. SRC solves a ℓ_1 -norm penalized least square problem and identifies the class label individually, by checking class by class. However, ℓ_1 -minimization makes SRC very computationally expensive, and recent research shows it lacks stability in face recognition [17, 18]. Furthermore, Zhang *et al.* point out the truth that SRC can produce interesting outcome is the contribution of collaborative representation [18], and they propose a new ℓ_2 regularized classification algorithm called *collaborative representation-based classification with regularized least square* (CRC-RLS), which adopts ℓ_2 -norm penalty rather than ℓ_1 regularizer in SRC. This modification leads to the simple ridge regression. However, when the number of collaborative data (training data) grows, calculating the coefficients by ridge regression becomes more computationally expensive, because a larger matrix inverse operation is involved.

Luckily, in a parallel process, focusing on high-dimensional sparse coding problem, Yu *et al.* empirically observe that sparse coding results tend to be local – nonzero coefficients are often assigned to bases nearby the encoded data. And they theoretically point out that under certain assumptions locality is more essential than sparsity [19]. To make full use of this local relation, they suggest a modification to sparse coding, called *local coordinate coding* (LCC), and based on their work, Wang *et al.* propose a practical method called *locality-constraint linear coding* (LLC) to fast implement LCC, and approximate LLC by utilizing *K-nearest neighbors* (KNN) ahead of time [20].

Combining the CRC-RLS and the idea of LLC, we develop a new classification mechanism, which adopts collaborative representation-based recipe regularized by ℓ_2 penalty and employs KNN search beforehand. Therefore, our classification approach is more stable to outliers as stated in [17], and much more efficient because only small-size ridge regression is involved even when the number of training samples is large.

3 The proposed Texture Classification Framework

Focusing on the four basic elements of a reliable texture classification system, in this section, we introduce our proposed framework in detail: local texture descriptors, overall texture image representation, measurement and classification mechanism. The overall flowchart of our approach is displayed by Figure 2. Notations used in this paper are embedded in Subsection 3.1.

3.1 Local Texture Descriptor

In our work, we use a single type of feature descriptor, the popular SIFT descriptor [8], which is extracted on a dense grid rather than at interest points and has been shown to yield superior classification performance in [10, 20, 15, 11]. Suppose there are T images from C classes and \mathcal{I}_c denotes the index of c^{th} class, and let t^{th} image be represented by a set of dense SIFT descriptors $\mathbf{x}_i^{(t)} \in \mathbb{R}^d$ ($d = 128$ for SIFT descriptor) at N locations identified with their indices $i = 1, \dots, N$. M regions of interest are defined on the image with \mathcal{N}_m denoting the set of locations/indices within region m , and $m \in \mathcal{L}_l$ means m^{th} region belongs to l^{th} level, *i.e.* \mathcal{L}_l indexes the regions in l^{th} level. Then we use all the dense SIFT descriptors to train a dictionary $\mathbf{D} \in \mathbb{R}^{d \times D}$, and employ the learned dictionary back to represent the dense SIFT descriptors into a sparse code vector, as the formulation

¹ It means an image is partitioned into 3×3 grid cells in the first level, 4×4 and 5×5 for the second and third level, respectively. Altogether $3 \times 3 + 4 \times 4 + 5 \times 5 = 50$ sub-images or regions or grid cells are formed.

below:

$$\begin{aligned}
(\mathbf{a}_i^{(t)}, \mathbf{D}) = \arg \min_{\mathbf{a}_i^{(t)}, \mathbf{D}} \sum_{t=1}^T \sum_{i=1}^N \{ \|\mathbf{x}_i^{(t)} - \mathbf{D}\mathbf{a}_i^{(t)}\|_2^2 + \lambda \|\mathbf{a}_i^{(t)}\|_1 \} \\
\text{s.t.} \quad \mathbf{d}_i^T \mathbf{d}_i \leq 1 \quad \text{for } i = 1, \dots, D \quad .
\end{aligned} \tag{1}$$

where $\mathbf{a}_i^{(t)} \in \mathbb{R}^D$ is the corresponding sparse code vector.

Each element a_k of the code vector \mathbf{a} indicates the local descriptor's response to the k^{th} visual word in the dictionary \mathbf{D} . We align all the SIFT descriptors belonging to region m as a matrix $\mathbf{X} \in \mathbb{R}^{d \times |\mathcal{N}_m|}$, then the corresponding code matrix $\mathbf{A} \in \mathbb{R}^{D \times |\mathcal{N}_m|}$ is obtained. Here we aggregate the local descriptors' responses across all the $|\mathcal{N}_m|$ locations of this region into an $|\mathcal{N}_m|$ -dimensional response vector \mathbf{a}_k^T (the k^{th} row of \mathbf{A}), in which each elements $a_{k,m}^T$ of \mathbf{a}_k^T represents the response of the local descriptor \mathbf{x}_m at the m^{th} location to the k^{th} visual word. After obtaining all the feature descriptors \mathbf{A} within a region, we can use a pooling operation to pool these feature descriptors into a single vector \mathbf{y} of fixed dimension, described in Subsection 3.1.2. Before feature pooling, we first address the relevant partition issues.

3.1.1 Partition Issues

Different from classical and commonly used SPM scheme, which is three or four level pyramid comprising pooling regions of $\{1 \times 1, 2 \times 2, 4 \times 4\}$ or $\{1 \times 1, 2 \times 2, 4 \times 4, 8 \times 8\}$ [12], we adopt a more flexible partition strategy and divide the original image into finer regions, *e.g.* $\{3 \times 3, 4 \times 4, 5 \times 5\}$ as Figure 1 shows.

Merely relying on this flexible partition fashion, through our observation, the proposed method can indeed capture local features in different scales and is resilient to local variance, such as translation, illumination and scale. But we go beyond by permitting different overlapping patterns at the same level of pyramid. Various overlapping patterns within a single level produce more regions when adopting the same partition pattern, *e.g.* a single level of 4×4 partition with 4 overlapping patterns will lead to $4 \times 4 \times 4 = 64$ regions, as displayed by row (c) in Figure 1, and accordingly 64 feature codes will be formed. More overlapping choices can produce more local texture features on multiple scales, and therefore these redundant local texture features can effectively alleviate the classification challenge caused by local variance.

This way of partition with multiple overlapping patterns prevents the statistical information or pooled feature codes of local regions from becoming too rigid or too flappy for texture discrimination, and in conjunction with our proposed classification mechanism described in Subsection 3.3, it will lead to state-of-the-art performance of texture classification in the experiments.

3.1.2 Feature Pooling

Feature pooling is essentially to map the response vectors within each region into a statistic value $f(\mathbf{a}_k^T)$ via some spatial pooling operation f . Among various pooling methods, such as average pooling, max pooling and some other pooling methods transiting from average to max [21], max pooling is inspired by the mechanism of the complex cells in the primary visual cortex and has been shown a powerful operation empirically and theoretically [10, 21, 11, 15]. In this paper, we also adopt max pooling for its translation-invariance in different level of partitions [22].

After obtaining code matrix \mathbf{A} of region m , we can pool the code vectors into one feature vector $\mathbf{y}_m \in \mathbb{R}^D$ to represent region m :

$$\begin{aligned}\mathbf{y}_m(\mathbf{A}) &= [f(\mathbf{a}_1^T), \dots, f(\mathbf{a}_k^T), \dots, f(\mathbf{a}_K^T)]^T \\ &= [\max_{i \in \mathcal{N}_m} a_{1,i}, \dots, \max_{i \in \mathcal{N}_m} a_{k,i}, \dots, \max_{i \in \mathcal{N}_m} a_{K,i}]^T\end{aligned}\quad (2)$$

Actually, no matter how the size of different regions differs, the pooled feature code is of the same dimension and well summarize the distribution of the SIFT feature descriptors in each region. This property enables us to adopt the flexible partition way and various overlapping patterns within the same level of partition, thereby producing redundant local texture features.

3.2 Texture Image Descriptor

As described in the previous subsection, we store all the pooled feature codes of one image to form a matrix $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_m, \dots, \mathbf{y}_M]$ as the new texture image representation. That is to say, regardless of region size and overlapping patterns, all the pooled feature vectors of regions are stored in an orderless way. This orderless storage, in conjunction with max pooling, enjoys translation and scale invariance. From Figure 1, it is not difficult to see the samples from the same class can represent one another by the statistic information that max pooling accumulates, which is local translation and scale invariant, therefore overall invariance property can be attained. We will see the benefit of this orderless storage from experiment in Section 4.

3.3 Measure and Classification

Actually, all the pooled feature vectors from regions of various levels of training images can be seen as redundant feature bases, or a feature pond, which can effectively represent pooled feature codes of a new image, and in this way, scale and translation invariance can be achieved. To fully take advantage of the benefit of orderless feature vector storage, we utilize a regularized least square (RLS) framework for the final classification. It is similar in appearance to sparse representation-based classification (SRC) [13], but essentially different.

In SRC, a vectorized test image \mathbf{z} is coded collaboratively over the dictionary of all T training samples $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_t, \dots, \mathbf{y}_T]$ under ℓ_1 -norm sparsity constraint, where \mathbf{y}_t is t^{th} vectorized training sample. For simplicity, SRC first calculate sparse coefficients by the formulation:

$$\mathbf{a} = \arg \min_{\mathbf{a}} \|\mathbf{z} - \mathbf{Y}\mathbf{a}\|_2^2 + \lambda \|\mathbf{a}\|_1 \quad (3)$$

Then, SRC classifies test image \mathbf{z} individually to determine which class \mathbf{z} should belong to. In other words, it calculates reconstruction error $r_c = \|\mathbf{z} - \mathbf{Y}_c \mathbf{a}_c\|_2$ for all the C classes, where \mathbf{Y}_c is formed by the columns indexed by \mathcal{I}_c and \mathbf{a}_c is formed in the similar way. Finally it selects $\hat{c} = \arg \min_c r_c$ as the predicted label.

Although SRC has shown interesting results in face recognition and has been widely studied in the community, researchers recently have found that, in SRC, ℓ_1 -norm penalty in Equation 3 actually makes the classification framework unstable [18, 17], as well as computationally very expensive. Zhang *et al.* point out the truth that SRC improves face recognition accuracy is the use of collaborative representation, but not ℓ_1 sparsity [18]. And they propose a collaborative representation-based classification framework with regularized least square (CRC-RLS) by solving a ridge regression formulation:

$$\mathbf{a} = \arg \min_{\mathbf{a}} \|\mathbf{z} - \mathbf{Y}\mathbf{a}\|_2^2 + \lambda \|\mathbf{a}\|_2^2, \quad (4)$$

Algorithm 1 Algorithm of LC-CRC

Input: feature descriptor matrix of testing image $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_M]$ and feature pond formed by all the training samples $\mathbf{Y} = [\mathbf{Y}_1, \dots, \mathbf{Y}_t, \dots, \mathbf{Y}_T]$, where $\mathbf{Y}_t = [\mathbf{y}_1, \dots, \mathbf{y}_M]$, parameter K for KNN search and λ for balancing ℓ_2 -norm penalty and least square fitting.

Output: predicted label of the test image.

- 1: Normalize the columns of \mathbf{Z} and \mathbf{Y} to have unit ℓ_2 -norm length;
- 2: **for** $m = 1, 2, \dots, M$ **do**
- 3: Use KNN within feature pond \mathbf{Y} , selecting K neighbors of \mathbf{z}_m to form matrix $\mathbf{Y}_{(K)} \in \mathbb{R}^{D \times K}$ with K indices \mathcal{H}_m ;
- 4: Code \mathbf{z}_m over $\mathbf{Y}_{(K)}$ by

$$\hat{\mathbf{a}}^m = (\mathbf{Y}_{(K)}^T \mathbf{Y}_{(K)} + \lambda \mathbf{I})^{-1} \mathbf{Y}_{(K)}^T \mathbf{z}_m;$$

- 5: Form MT -dimensional vector \mathbf{a}^m where elements of \mathcal{H}_m locations are embedded with $\hat{\mathbf{a}}^m$ and zeros elsewhere;
- 6: **end for**
- 7: Compute the reconstruction error for each class:

$$r_c = \sum_{l=1}^L \left\{ \min_{m \in \mathcal{L}_l} \|\mathbf{y}_m - \mathbf{Y}_c \hat{\mathbf{a}}_c^m\|_2 \right\};$$

- 8: Output the identity of test image \mathbf{Y} as:

$$\text{identity}(\mathbf{Y}) = \arg \min_c (r_c).$$

Following the rest part of SRC, CRC-RLS achieves very competitive classification results but with significantly less complexity than SRC.

However, when the number T of training samples grows, calculating the coefficients by ridge regression $\mathbf{a} = (\mathbf{Y}^T \mathbf{Y} + \lambda \mathbf{I})^{-1} \mathbf{Y}^T \mathbf{z}$ becomes more computationally expensive, because inverse operation on a larger matrix of size $T \times T$ is involved. To circumvent this problem, we borrow the idea of LLC [20] described in Section 2 by applying KNN search among the feature pond before solving the ridge regression — choosing K nearest neighbors to form $\mathbf{Y}_{(K)} \in \mathbb{R}^{D \times K}$ with indices $\mathcal{H}_{(K)}$, and representing the testing image by solving a much lower-complexity ridge regression: $\hat{\mathbf{a}} = (\mathbf{Y}_{(K)}^T \mathbf{Y}_{(K)} + \lambda \mathbf{I})^{-1} \mathbf{Y}_{(K)}^T \mathbf{z}$. After this, an overall coefficient vector $\mathbf{a} \in \mathbb{R}^T$ is formed by embedding the elements of $\hat{\mathbf{a}} \in \mathbb{R}^K$ in $\mathcal{H}_{(K)}$ locations of \mathbf{a} and zeros elsewhere. The final classification follows SRC, and Algorithm 1 shows the whole classification algorithm².

4 Experiment

We evaluate the performance of the proposed texture classification framework on three public datasets: Brodatz dataset [23], KTH-TIPS dataset [24] and UMD texture database [1].

The Brodatz dataset is a well-known benchmark database for evaluating texture recognition algorithms. It contains 111 different texture classes. For each class, it is represented by only one sample, which is then divided into 9 sub-images non-overlappingly to form the database. Thus, there are

² Because of one texture image is represented by a descriptor matrix as Subsection 3.2 introduces, in the algorithm, each column of descriptor matrix should be treated individually, and the final reconstruction error is to accumulate over L columns (each column denote a pooled feature code of a specific region) — the summation of the smallest error of each level. Empirically, we find the using of smallest error for classification brings out better performance than that of the reconstructive errors of all the codes.

999 images altogether with resolution of 215x215. Although this dataset lacks interclass variations, Lazebnik *et al.* point out that this database is a challenging platform for testing the discriminative power of texture descriptors, thanks to its variety of scales and geometric patterns [4]. The KTH-TIPS textures dataset contains ten texture classes. Images are captured at nine scales spanning two octaves (relative scale changes from 0.5 to 2), viewed under three different illumination directions and three different poses, thus giving a total of 9 images per scale, and 81 images per material class. Some sample images are shown in Figure 5, and we can easily see the scaling and illumination changes increase the intra-class variability and makes this database especially difficult for classification task. UMD texture database is composed of 25 different texture classes, 40 samples for each, and all images are grayscale of 1280x960 pixels (1000 samples altogether). The textures are acquired under strong viewpoint and scale changes, arbitrary rotations, and significant contrast differences, even including textures with nonrigid deformation. Figure 1 displays some sample images from this database.

4.1 Configurations and Implementation

Dictionary learning and sparse coding. As a dictionary learning problem, Equation 1 is convex in $\mathbf{a}_i^{(t)}$ with fixed \mathbf{D} and vice versa, but not convex simultaneously for both of them. The conventional way for such problem is to solve it iteratively by alternately optimizing over \mathbf{D} or $\mathbf{a}_i^{(t)}$'s while fixing the other. Fixing \mathbf{D} , the optimization can be solved by optimizing over each coefficient $\mathbf{a}_i^{(t)}$ individually:

$$\min_{\mathbf{a}_i^{(t)}} \|\mathbf{x}_i^{(t)} - \mathbf{D}\mathbf{a}_i^{(t)}\|_2^2 + \lambda \|\mathbf{a}_i^{(t)}\|_1$$

This is essentially a linear regression problem with ℓ_1 -norm regularization on the coefficients, *i.e.* Lasso in the Statistical literature. In our work, we solve this optimization by a very efficient algorithm called *feature-sign search* [25]. Fixing all the $\mathbf{a}_i^{(t)}$'s, the problem is reduced to a least square problem with quadratic constraints:

$$\min_{\mathbf{D}} \|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2, \quad \text{s.t. } \|\mathbf{d}_i\|_2 \leq 1 \text{ for } i = 1, \dots, D.$$

The optimization can be done efficiently by the Lagrange dual as used in [25]. Throughout this paper, parameter λ is set 0.1. In our experiment, the dictionaries learned contain 1500 visual words for Brodatz and KTH-TIPS dataset, and 3000 visual words for UMD dataset.

Partition strategy and overlapping patterns. For the experiments, we partition all the texture images into 4 levels ($2 \times 2, 3 \times 3, 4 \times 4, 5 \times 5$) over Brodatz dataset, 3 levels ($6 \times 6, 7 \times 7, 8 \times 8$) for KTH-TIPS dataset, and 4 levels ($3 \times 3, 4 \times 4, 5 \times 5, 6 \times 6$) for UMD texture database. Furthermore, over each partition level, we admit various overlapping patterns. Actually, we empirically find the partition strategy of each three datasets produces satisfactory results.

LC-CRC framework for classification. In our proposed LC-CRC classification framework, there is a parameter λ (different from the one of dictionary learning in Equation 1) to make the solution of the least square problem Equation 4 stable. Through empirical observations, we find that the experimental results are not sensitive to the choice of λ if a small value is assigned which is less than 0.01, and thus we set λ as 0.001 through out our work. Moreover, parameter K of KNN algorithm ought to be specified, and we set $K = 100$ when the number of training samples per class is small, *e.g.* only 1 or 2 samples of each class are available for training, and $K = 300$ when more training samples per class are available.

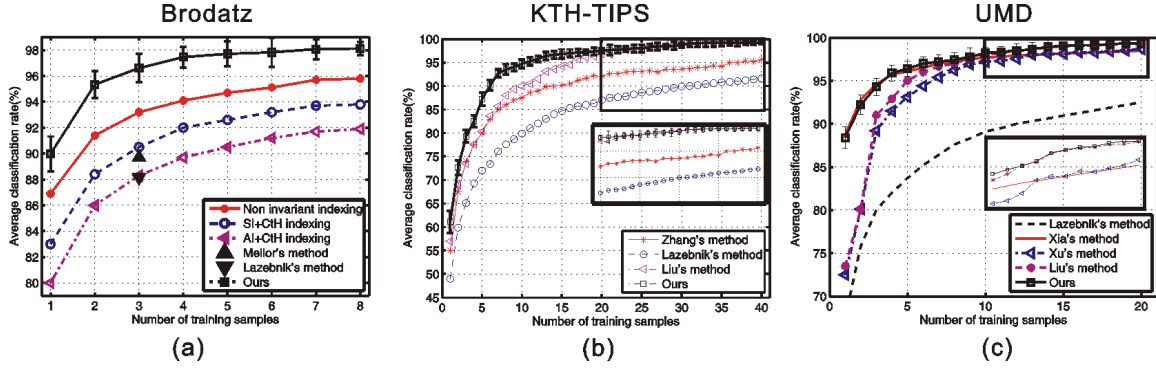


Figure 3: Classification rate vs. number of training samples on the three datasets.

4.2 Brodatz dataset

Figure 3 (a) shows the classification results obtained on the Brodatz database. Following Lazebnik *et al.* [26] and Mellor *et al.* [27], classification rates are estimated by averaging the results on randomly selected training sets, and 10 trials are performed in our experiments.

SI+CtH indexing and AI+CtH indexing are two shape-based invariant texture features from the work of Xia *et al.* [28]. Here SI proposed by Xia *et al.* is a kind of feature that is invariant to (local) similarity transforms, and AI means (locally) affine invariant features. Both of them are made of several histograms, such as scale ratio histogram, elongation histogram and compactness histogram. Moreover, CtH is contrast histogram computed by scanning all pixels of a local adaptive neighborhood, which is robust to geometrical distortions of the textures [28]. Due to that the samples in this dataset are created by cutting each texture of the Brodatz database into pieces, as a consequence, the resulting dataset lacks of viewpoint and scale changes. For this reason, Xia *et al.* also adopt a well chosen non-invariant indexing scheme (Non-invariant indexing in Figure 3 (a)) and it shows better classification result. Despite multiple histograms in [28], our framework only employs one kind of feature descriptor (SIFT), and it achieves state-of-the-art performance. Note that when 3 samples per class are used for training, our approach achieves 96.61%. This outcome is higher than 95.9% achieved by the method of Zhang *et al.*, based on the method of Lazebnik *et al.* [26] by employing three types of descriptors (SPIN, RIFT and SIFT) [5], and is comparable with 97.16% (the highest classification rate on Brodatz dataset to the best of our knowledge) attained by the method of Liu *et al.* by using sorted random projections plus several kernel SVMs [3]. However, it is worth noticing that when only one images of each class is used for training, our approach achieves 90% accuracy, which is significantly higher than the other methods. This verifies that our approach indeed can extract large amount reliable features of each type of textures, even when only a few sample images are available for training.

4.3 KTH-TIPS texture database

Following Zhang *et al.* [5], we vary the number of training images and record classification accuracy, as Figure 3 (b) shows. Note that all images are converted to grey scale in our approach and no use of color information is made whatsoever. Three methods are used for comparison, and the results of these methods are taken directly from the original publications or quoted from the recent comparative study of Zhang *et al.* [5]. In [26], Lazebnik *et al.* first characterize the texture using Harris-affine corners and Laplacian-affine blobs with two descriptors (SPIN and RIFT), and employ nearest

aluminium foil	98.59	0.00	0.00	0.00	1.41	0.00	0.00	0.00	0.00	0.00
brown bread	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
corduroy	0.00	1.41	91.55	7.04	0.00	0.00	0.00	0.00	0.00	0.00
cotton	0.00	0.00	1.41	90.14	0.00	2.82	1.41	4.23	0.00	0.00
cracker	2.82	2.82	0.00	0.00	94.37	0.00	0.00	0.00	0.00	0.00
linen	0.00	0.00	0.00	11.27	2.82	81.69	1.41	2.82	0.00	0.00
orange peel	1.41	0.00	1.41	0.00	0.00	0.00	97.18	0.00	0.00	0.00
sandpaper	0.00	4.23	0.00	0.00	2.82	0.00	0.00	90.14	2.82	0.00
sponge	0.00	0.00	0.00	0.00	1.41	0.00	0.00	0.00	98.59	0.00
styrofoam	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.41	0.00	98.59

Figure 4: Performance on KTH-TIPS texture database, confusion matrix for classification of 10 different textures. 10 images per class are randomly selected for training, and the rest for testing. The number at row R and column C is the proportion of R class which is classified as C class. For example, 9.86% of the linen images are misclassified as cotton class. The average accuracy is 94.23%.

neighbor classifier. Their method achieves 91.3% accuracy when 41 samples of each class are used for training. Under the same configuration, the method of Zhang et al, introduced in Subsection 4.2, achieves 96.1%, and the approach of Liu *et al.* achieves 99.29% (the highest classification rate on KTH-TIPS dataset to the best of our knowledge) in [3]. And our method achieves 99.32% under the condition that 41 samples per material are used for training, which exceeds the best one (99.29%).

It is worth noting that our approach achieves $(94.1 \pm 0.92)\%$ when only 10 images of each class are randomly selected for training, which is significantly higher than the others. Figure 4 displays one confusion matrix under this condition. From the confusion matrix, we can see the misclassifications mainly concentrate on four materials: corduroy, cotton, linen and sandpaper. Figure 5 shows some samples of the four types of materials, and it can be easily seen that under different scale of different materials, they are very similar and this phenomenon results misclassification within these material types.

4.4 UMD texture database

The UMD texture database contains images of larger arbitrary rotation, larger scale variation and more significant viewpoint than the previous two datasets. Therefore, it is more challenging for classification.

Figure 3 (c) shows the classification rate vs. the number of training samples on UMD dataset. Xia’s method denotes the SI+CtH indexing method as described in Subsection 4.2 in conjunction with geodesic distance, which considers textures as points lying on some intrinsic manifold and yields clear improvement in their method. Xu’s method is based on a combination of wavelet transform and multifractal analysis. Liu’s method is introduced previously in Subsection 4.2. We can see when only a few samples of each class are available for training, our method is comparable to Xia’s method, which achieves the best performance on this database under small amount of training samples. While the number of training images of each class is increasing, Liu’s method obtains better results. Still, Under this condition, our method achieves comparable outcome with Liu’s method. When 20 sample

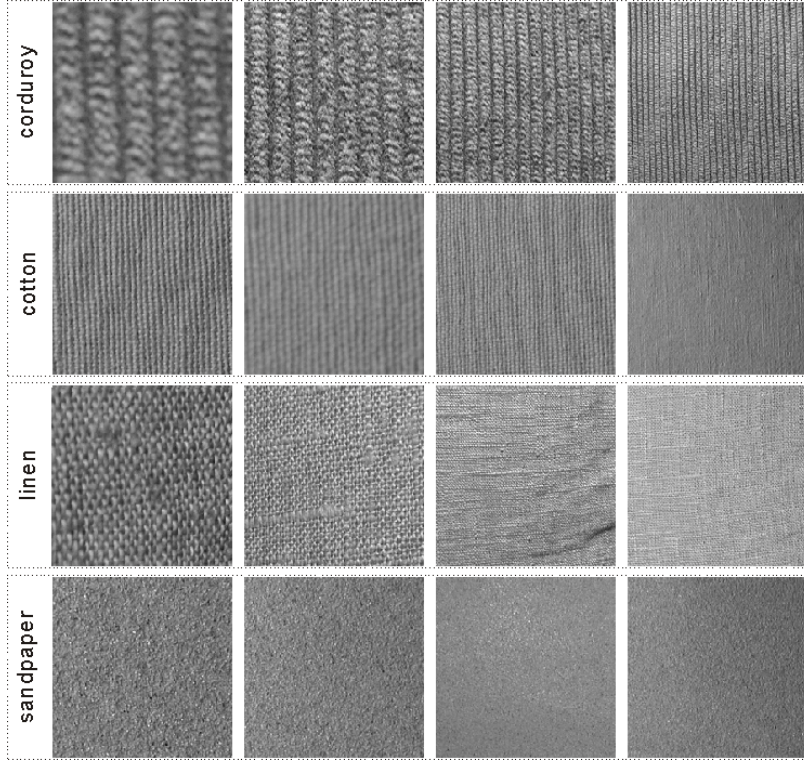


Figure 5: Similar texture pattern on KTH-TIPS database. Four different texture types are displayed here, it is easy to see that some images of the four textures are very similar with various scales. This phenomenon largely leads to misclassification on this dataset.

images per class are randomly selected for training, 98.6% classification accuracy is achieved by Xu’s method and 99.30% by Liu’s method (the best one ever reported on this database). And our method achieves $(99.32 \pm 0.35)\%$ classification accuracy, which is slightly better than Liu’s result.

From this experiment on UMD dataset, we can see our proposed texture classification approach can extract reliable texture features while only a few training sample images are available and leads to significantly better results. While the number of training sample grows, our method can still achieve state-of-the-art performance compared with other methods. In the consideration of the complex texture sample images from UMD dataset, it is easy to see that our method achieves invariance to local rotation variation, scale changes, translation, changes of illumination directions and significant viewpoint.

5 Conclusion and Future Work

In this paper, focusing on texture classification task, we introduce a novel and highly effective scheme for robust texture classification, which is invariant to scale differences, translation, significant viewpoint changes and local rotation. Inspired by SPM framework, we first develop a multi-level descriptor to describe local texture features, allowing different level of partitions and various overlapping patterns within each level of partition. From experiments, we see this flexible descriptor can better capture the local features of each kind of texture, and even when only a few samples of each class

are available for training, our method still achieves very high accuracy. Accordingly, we propose an efficient classification mechanism, which is based on collaborative representation with locality constraint, called LC-CRC. It first search relatively a few neighbors from the feature pond by KNN algorithm, and then use them to represent the target through solving a simple least square fitting problem with ℓ_2 -norm regularization. To evaluate our texture classification framework, we conduct several experiments on three well-known texture datasets and the outcome is very competitive and even outperforms several state-of-the-art methods.

Actually, LC-CRC classification framework treat the feature pond as another dictionary, which is used to represent the pooled feature codes of testing images. This spirit of hierarchical sparse coding has been already explored by Yu *et al.* in [29] for object recognition, but there remains interesting extensions and confirmations, and one of our future work is to provide some insights of multi-layer dictionary learning for image classification. Moreover, our work provides a new application of SPM, and we expect some other applications based on SPM and its variants.

Acknowledgements

This work is supported by by 973 Program (Project No.2010CB327905) and Natural Science Foundations of China (No.61071218).

References

- [1] Y. Xu, H. Ji, and C. Fermuller, “Viewpoint invariant texture description using fractal analysis,” *IJCV*, 2009.
- [2] Y. Xu, X. Yang, H. Ling, and H. Ji, “A new texture descriptor using multifractal analysis in multi-orientation wavelet pyramid,” *CVPR*, 2010.
- [3] L. Liu, P. Fieguth, G. Kuang, and H. Zha, “Sorted random projections for robust texture classification,” *ICCV*, 2011.
- [4] S. Lazebnik, C. Schmid, and J. Ponce, “A sparse texture representation using local affine regions,” *PAMI*, 2005.
- [5] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, “Local features and kernels for classification of texture and object categories: A comprehensive study,” *IJCV*, 2007.
- [6] M. Crosier and L. D. Griffin, “use basic image features for texture classification,” *IJCV*, 2010.
- [7] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognition natural scene categories,” *CVPR*, 2006.
- [8] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *IJCV*, 2004.
- [9] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” *CVPR*, 2005.
- [10] J. Yang, K. Yu, Y. Gong, and T. Huang, “Linear spatial pyramid matching using sparse coding for image classification,” *CVPR*, 2009.
- [11] Y.-L. Boureau, N. L. Roux, F. Bach, J. Ponce, and Y. LeCun, “Ask the locals: multi-way local pooling for image recognition,” *ICCV*, 2011.
- [12] T. Harada, Y. Ushiku, Y. Yamashita, and Y. Kuniyoshi, “Discriminative spatial pyramid,” *CVPR*, 2011.

- [13] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *PAMI*, 2008.
- [14] K. Grauman and T. Darrell, “The pyramid match kernel: Discriminative classification with sets of image feature,” *ICCV*, 2005.
- [15] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce, “Learning mid-level features for recognition,” *CVPR*, 2010.
- [16] Y. Meyer, “Workshop: An interdisciplinary approach to textures and natural images processing,” *Institut Henri Poincaré, Paris*, 2007.
- [17] Q. Shi, A. Eriksson, A. van den Hengel, and C. Shen, “Is face recognition really a compressive sensing problem?,” *CVPR*, 2011.
- [18] L. Zhang, M. Yang, and X. Feng, “Sparse representation or collaborative representation: Which helps face recognition?,” *ICCV*, 2011.
- [19] K. Yu, T. Zhang, and Y. Gong, “Nonlinear learning using local coordinate coding,” *NIPS*, 2008.
- [20] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, “Learning locality-constrained linear coding for image classification,” *CVPR*, 2010.
- [21] Y.-L. Boureau, J. Ponce, and Y. Lecun, “A theoretical analysis of feature pooling in visual recognition,” *ICML*, 2010.
- [22] J. Yang, K. Yu, and T. Huang, “Supervised translation-invariant sparse coding,” *CVPR*, 2010.
- [23] P. Brodatz, “Textures: A photographic album for artists and designers,” *New York: Dover*, 1966.
- [24] E. Hayman, B. Caputo, M. Fritz, and J.-O. Eklundh, “On the significance of real-world conditions for material classification,” *ECCV*, 2004.
- [25] H. Lee, A. Battle, R. Raina, and A. Y. Ng, “Efficient sparse coding algorithms,” *NIPS*, 2007.
- [26] J. P. Svetlana Lazebnik, Cordelia Schmid, “A sparse texture representation using local affine regions,” *PAMI*, 2005.
- [27] M. Mellor, B.-W. Hong, and M. Brady, “Locally rotation, contrast, and scale invariant descriptors for texture analysis,” *TPAMI*, 2008.
- [28] G.-S. Xia, J. Delon, and Y. Gousseau, “Shape-based invariant texture indexing,” *IJCV*, 2010.
- [29] K. Yu, Y. Lin, and J. Lafferty, “Learning image representations from the pixel level via hierarchical sparse coding,” *CVPR*, 2011.